

# Distributed association rule mining and summarization for Diabetes Mellitus and Its Co-Morbid Risk Prediction strategy using FUZZY Classifier

Dhivya Selvaraj, Mrs.Merlin Mercy

**Abstract**— Diabetes is a life-threatening issue in modern health care domain. With the use of data mining techniques, diabetes factors and co morbid risk conditions associated with diabetes has found. In order to stifle the evolution of diabetes mellitus, applies distributed association rule mining and summarization techniques to electronic medical records. This helps to discover set of risk factors and co morbid conditions in distributed medical dataset using frequent itemset mining. In general, association rule mining (ARM) generates bulky volume of data sets which need to summarize certain rules over medical record. This encompasses a novel approach to find the common factors which lead to high risks of diabetes and co morbid conditions associated with diabetes. This performs both association rule mining and association rule summarization techniques with improved classification algorithms. Existing systems aim to apply association rule mining to electronic medical records to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes. Given the high dimensionality of EMRs (Electronic Medical Records), association rule mining generates a very large set of rules which we need to summarize for easy clinical use. The existing system reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding the diabetes risk prediction.

**Index Terms**— Diabetes, Association, Co-Morbid, Prediction, Split and Merge ,FUZZY

## I. INTRODUCTION

Diabetes is a group of metabolic diseases caused by hyperglycemia this is because of defects in insulin secretion, insulin action and both. Next stage chronic hyperglycemia of diabetes is associated with long term damage, dysfunction, and failure of different organs of body, especially the eyes, kidneys, nerves, heart, and blood vessels. This deficiency leads to destruction of the b-cells of the pancreas with consequent insulin deficiency to abnormalities that result in resistance to insulin action and reaction process. The basis of the abnormalities found in carbohydrate, fat, and protein metabolism in diabetes is deficient action of insulin on target tissues. Deficiency of insulin results from inadequate insulin secretion otherwise diminished tissue responses to insulin at the complex pathways of hormonal reaction in the body. Improper insulin secretion and defects in insulin action that is

frequently coexist in the same patient and it is often unclear which abnormality is the primary cause of the hyperglycemia. The symptoms of marked hyperglycemia include which includes polyuria, polydipsia, weight loss, sometimes with polyphagia, and blurred vision impairment. The growth and susceptibility to certain infections may also accompany chronic hyperglycemia in so many patients. Life-threatening consequences of uncontrolled diabetes are hyperglycemia with keto acidosis and the non ketotic hyperosmolar syndrome. Long-term complications of diabetes causes retinopathy with potential vision loss, nephropathy leading to renal failure, peripheral neuropathy with foot ulcers, amputations, Charcot joints, and autonomic neuropathy causing gastrointestinal, genitourinary, and cardiovascular symptoms and sexual dysfunction. Patients with diabetes have an increased incidence of atherosclerotic cardiovascular, peripheral arterial and cerebrovascular disease. Hypertension and abnormalities of lipoprotein metabolism are often found in people with diabetes. The vast majority of cases of diabetes fall into two broad etiopathogenetic categories. In one category, type 1 diabetes, the cause is an absolute deficiency of insulin secretion. Individuals at increased risk of developing this type of diabetes can often be identified by serological evidence of an autoimmune pathologic process. In the other, much more prevalent category, type 2 diabetes, the cause is a combination of resistance to insulin action and an inadequate compensatory insulin secretory response. In the latter category, a degree of hyperglycemia sufficient to cause pathologic and functional changes in various target tissues, but without clinical symptoms, may be present for a long period of time before diabetes is detected. During this asymptomatic period, it is possible to demonstrate an abnormality in carbohydrate metabolism by measurement of plasma glucose in the fasting state or after a challenge with an oral glucose load.

## II. CLASSIFICATION OF DIABETES MELLITUS AND OTHER CATEGORIES OF GLUCOSE REGULATION:

Assigning a type of diabetes to an individual often depends on the circumstances present at the time of diagnosis, and many diabetic individuals do not easily fit into a single class. For example, a person with gestational diabetes mellitus may continue to be hyperglycemic after delivery and may be determined to have, in fact, type 2 diabetes. Alternatively, a person who acquires diabetes because of large doses of exogenous steroids may become normoglycemic once the glucocorticoids are discontinued, but then may develop diabetes many years later after recurrent episodes of pancreatitis. Another example would be a person treated with

Dhivya Selvaraj, ME PG Scholar, Department of Computer Science, Sri Krishna college of Technology, Coimbatore, Tamil Nadu.

Mrs.Merlin Mercy, Asst professor, Department of Computer Science, Sri Krishna college of Technology, Coimbatore, Tamil Nadu.

thyroids who develops diabetes years later. Because thiazides in themselves seldom cause severe hyperglycemia, such individuals probably have type 2 diabetes that is exacerbated by the drug. Thus, for the clinician and patient, it is less important to label the particular type of diabetes than it is to understand the pathogenesis of the hyperglycemia and to treat it effectively.

### 2.1 ABNORMALITY DIAGNOSIS:

In data mining, abnormality detection is the search for data items in a dataset which do not conform to an expected pattern. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. Anomalies are also referred to as outliers. A good definition of an abnormality is as follows “an abnormality is an observation that deviates so much from other observations as to arouse suspicions that it was caused by a different mechanism”. Distance-based measures have been used in algorithms to delineate outliers or abnormal records from normal records.

Abnormality detection is the process of identifying abnormal pattern from set of objects. Abnormality detection aims to identify a small group of instances which deviate remarkably from the existing data. A well-known definition of “abnormality” is given “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism,”.

- Abnormality detection refers to the problem of finding patterns in data that do not conform to expected normal behavior.
- These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains.

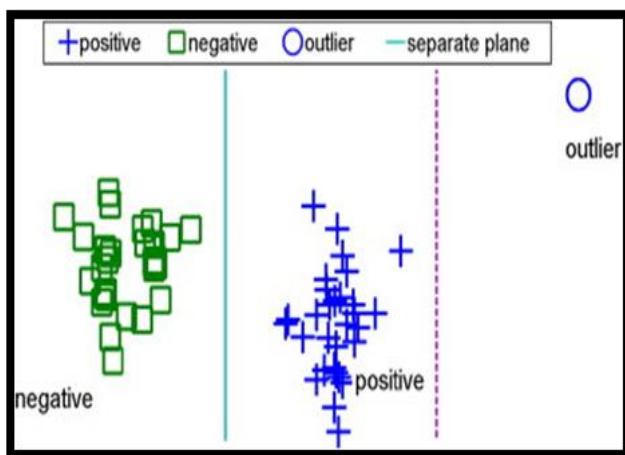


Fig: 2 abnormality detection

### III. RELATED WORK

Existing systems aim to apply association rule mining to electronic medical records to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes.

Given the high dimensionality of EMRs (Electronic Medical Records), association rule mining generates a very large set of rules which we need to summarize for easy clinical use. The existing system reviewed four association rule set

summarization techniques and conducted a comparative evaluation to provide guidance regarding the diabetes risk prediction.

The Disadvantages include:

- Less accuracy
- Need more training data
- Processing overhead
- Only predicts the diabetes risk failed to perform the co-morbid conditions and its risk.

### IV. LITERATURE SURVEY

Data mining has been participated an imperative role in the intelligent medical systems which is stated in paper [3][8][9]. The associations of disorders and the real causes of the disorders and the effects of symptoms that are impulsively seen in patients can be evaluated by the users via data mining techniques. In the application of health domain, Bulky databases can be applied as the input data to the system to find the association between attributes. The effects of associations have not been evaluated adequately in the literature. This have been explored the relationships of hidden knowledge placed among the large medical databases. This has been searched relevant attributes by means of finding frequent items using candidate generation.

Learning of the risk factors associated with diabetes helps health care professionals to identify patients at high risk of having diabetes disease. Statistical analysis and data mining techniques [10] helps to healthcare professionals in the diagnosis of heart oriented diseases. Such analysis has identified the disorders of the heart and blood vessels, using statistical values, and this includes cerebrovascular disease known as stroke, coronary heart disease also known as heart attacks, raised blood pressure [hypertension], heart failure, rheumatic heart disease, peripheral artery disease and congenital heart disease.

In paper [11] presented an efficient approach for the prediction of heart attack risk levels from the heart disease dataset using clustering techniques. Initially the heart disease dataset is clustered using the K-means clustering algorithm, which will extract the attributes and data relevant to heart attack from the dataset. This allows the dataset to be portioned into k fragments. This approach mines the frequent patterns subsequently from the extracted data related to heart disease. This used MAFIA a maximal frequent Item set algorithm, which is a machine learning algorithms trained with selected significant patterns. This basically predicts the heart attack. Additionally some technique from [12] resolves the prediction accuracy oriented issues. The approach utilizes the ID3 algorithm as a training algorithm. The results showed that the designed prediction system is capable of predicting the heart attack effectively. But the prediction of diabetes is slightly different from the above.

A study on the prediction of heart attack risk levels from the heart disease database with the use of bayes algorithms has conducted in [13]. This utilized the basic data mining classification techniques with 11 important attributes. Mainly that is concentrated the bagging technique. From the results of [13] bagging technique is accurate and capable than the J48 and Bayesian classification algorithms for heart attack prediction.

In a predictive model, scores will be calculated to estimate the risk of diabetes, so there is a need of diabetes index. The need of diabetes index has been recognized in [2][15], this conducted a survey regarding the diabetes risk factors. They found that most indices were additive in nature and none of the surveyed indices have taken interactions among the risk factors into account.

Paper [14] used association rule mining to systematically explore associations of diagnosis codes. The resulting association rules do not constitute a diabetes index because the study does not designate a particular outcome of interest and they do not assess or predict the risk of diabetes in patients, but they discovered some significant associations between diagnosis codes.

Understanding Random forests [17] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost [18], but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Random forests are an effective tool in prediction [17]. Because of the Law of Large Numbers they do not over fit. Injecting the right kind of randomness makes them accurate classifiers and regressors. Furthermore, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation. For a while, the conventional thinking was that forests could not compete with arcing type algorithms in terms of accuracy. The results dispel this belief, but lead to interesting questions. Boosting and arcing algorithms have the ability to reduce bias as well as variance [20].

## V. PROPOSED SYSTEM:

The propose technique named as SAM (Split and Merge), which is based on fast distributed quantitative association rule mining and rule filtering for prediction co morbid conditions associated with diabetes. In the field of medical domain, the prediction of diabetes and its Co-Morbid in earlier stage is important. We propose a set of methods to perform the Co-Morbid prediction. This chapter specifies the process included in the proposed system.

### 3.1 SAM Strategy

#### 3.1.1 Association Rule in Health care

Data mining technique such as Association rule mining is applied to discover patterns or associations encoded in the data. Association rule is in the form of  $A \rightarrow B$  where  $A$  is the antecedent and  $B$  is the consequent and  $A$  and  $B$  are sets of predicates. The association rule is based on concepts of support and confidence. The support is the probability of a

transaction/event in the database containing both the antecedent and the consequent and the confidence is the probability that a record that contains the antecedent also

contains the consequent. If  $I = \{i_1, i_2, \dots, i_n\}$  is a set of items, a transaction  $T$  is a subset of  $I$ , and dataset  $D$  is set of transaction. Association rule then means finding rules in the form of

$$R \Rightarrow i[S, C] \quad (3.3)$$

where  $R \subseteq I$  and  $i \in I$ ,  $S$  is the support and  $C$  is the confidence. The support,  $support D(X)$  of an item  $X$  in the dataset can be defined as

$$SupportD(X) = \frac{countD(X)}{|D|} \quad (3.4)$$

where  $countD(X)$  is the number of transactions in  $D$  containing  $X$ . The user specifies a minimum support ( $min\_sup$ ) and confidence value ( $min\_conf$ ). An itemset is said to be frequent if its support is greater than the  $min\_sup$  value specified. Number of algorithms are proposed for discovering association rules from large database (Agrawal et al 1994; Han, et al 2000; Berzal et al 2001).

The *apriori algorithm* (Agrawal et al 1994) is on the most popularly used algorithms for discovering association rules. The algorithm first discovers all frequent itemsets  $I_F \subseteq I$  which has a value of support equal to or greater than  $min\_sup$ . The algorithm merges all the frequent itemsets

until no more  $I_F$  are found. On generation of the frequent itemsets, it is split in any possible way into a rule antecedent  $R \subseteq I$  and a rule consequent  $i \in I$  such that  $R \cup i = I_F$  and  $R \cap i = \phi$ . The confidence is calculated for each rule candidate and the rule is output if the confidence is above  $min\_conf$ .

Health attributes are the data related to disease diagnosis. In co-morbid association rule mining, it seeks to find associations among transactions that are encoded explicitly in a database, association rule mining seeks to find patterns in spatial relationships that are typically not encoded in a database but are rather embedded within the health care framework. These rules must be extracted from the data prior to the actual association rule mining.

Association mining rule is applied on the data collected to discover patterns. Every node mines patterns in the following form:

$$A_1 \wedge \dots \wedge A_m \Rightarrow E[S, C] \quad (3.5)$$

where event  $E$  occurs at node  $n$  with support  $S$  and confidence  $C$  given that antecedents  $A_i$  holds true. Antecedents are in the form of

$$A_i = (E_i, D_i, T_i, N_i) \quad (3.6)$$

Every iteration selects a subset of discovered patterns to the system, thus, reducing the computation overhead. Association rules describe the implication of one or a set of features by another set of features in health care databases.

#### 3.1.2 SAM algorithm

In this proposed system the SAM algorithm which is an extension of Apriori is proposed for generating association rule is as follows:

1. Initiate the process by uploading the dataset D
2. Apply SAM algorithm to find the frequent itemsets A with the minimum support.
3. Set  $R = \Phi$  where R is the rule set, which contains the association rule.
4. Represent each frequent item set of A as quantity data using the combination of representation.
5. Select the two members from the frequent item set and predict the co morbid risk by the SAM algorithm.
6. The next iteration is applying the crossover and mutation on the selected rule set to generate the final priority association rules.
7. Find the fitness function for each rule  $x \rightarrow y$  and check the following condition.
8. If (fitness function > min confidence)
9. Set  $R = R \cup \{x \rightarrow y\}$
10. If the desired number of generations is not completed, then go to Step 5.

The above steps explain the process of HARS. The algorithm terminates the execution when the condition is met. It also terminates execution when the total number of generations specified by the user is reached. The support of an association pattern is the percentage of task-relevant data tuples for which the pattern is true.

SAM algorithm is used to discover the frequent data item sets and classifying summarized and summarized data sets.

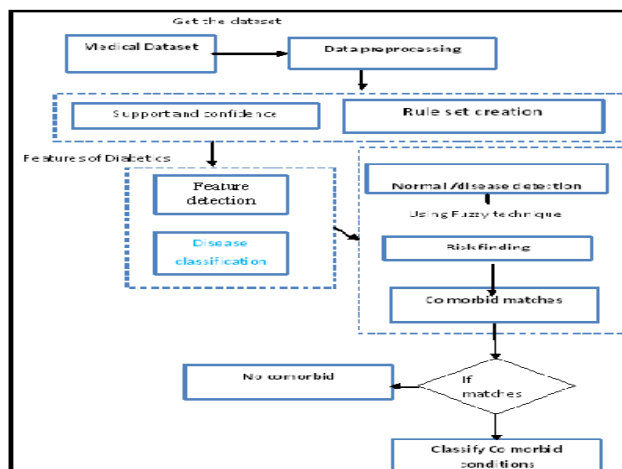


Fig 3.1 Block diagram of proposed system

### 3.2 FUZZY classifier

To classify the patient with diabetes.

1. In fuzzy grouping of data every point has a degree belonging to as in fuzzy logic rather than belonging completely to just one same cluster.
2. These co morbid conditions are the diseases associated with the diabetes that are realized by the diabetes patients.
3. These subtype diseases are all analyzed in the proposed system to make the application very efficient to use.

## VI. PROBLEM FORMALIZATION

The main contribution is a comparative evaluation of these extended summarization techniques that provides guidance to practitioners in selecting an appropriate algorithm for a similar problem.

To handle the existing problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Here, a weighted Association rule-based clinical decision support system is presented for the diagnosis of heart disease and diabetes, automatically obtaining knowledge from the patient's clinical data.

The proposed system for the risk prediction of heart and diabetes patients consists of two phases:

- (1) Automated approach for the generation of distributed association rules and summarization
- (2) Developing a Genetic-based decision support system to improve prediction accuracy.
- (3) Implementing a new association rule summarization algorithm named as SAM (Split and Merge algorithm)

In the first phase, we have used the attribute selection and attribute weightage method to obtain the weighted association rules. Then, the proposed system is constructed in accordance with the weighted rules and chosen attributes. The experimentation is carried out on the proposed system using the datasets obtained from the UCI repository.

## VII. EXPERIMENTAL ANALYSIS

### 5.1 Accuracy Comparison Chart

The below chart describes the accuracy comparison for our proposed system. This has been measured with the following formula.

BUS algorithm and SAM technique is compared and the results are classified based on

- Time and storage complexity
- Less overhead
- Effective risk identification
- Comorbid conditions
- Good accuracy
- Support and confidence level
- Other risk factors and diseases
- Diabetic and non diabetic
- Diabetic concerned risk factors

Rules	Accuracy(BUS)	Accuracy(HARS)
0.1	79	80
0.2	81	82

0.3	82	83.5
0.4	83	85
0.5	84.2	86.7
0.6	85.1	89
0.7	87	92
0.8	89	95
0.9	91	97

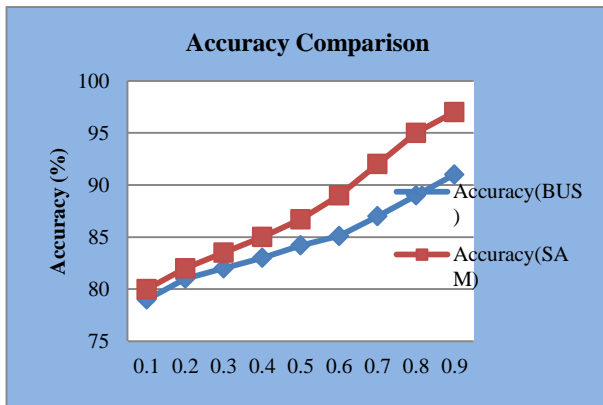


Fig 4.1: Performance comparison of proposed SAM with existing BUS approaches based on prediction efficiency.

## 5.2 Prediction of Co-Morbid conditions Accuracy Comparison Chart:

Iterations	SAM
1	80
2	82
3	89.5
4	93
5	94.7
6	95
7	96.8

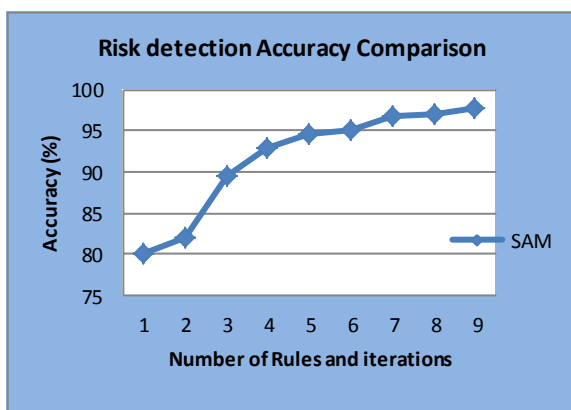


Fig 15: Performance comparison of proposed SAM with existing BUS approaches based on prediction accuracy

## VIII. CONCLUSION

The study proposed a new association summarization and Co-Morbid prediction scheme for diabetes and co-morbid conditions associated with diabetes. The system studied the main two problems in the literature, which are prediction accuracy and prediction error. The study overcomes the

above two problem by applying the effective hybrid association rule summarization with split and merge algorithm. The **SAM** (Split and Merge) represents with the effective rule specification criteria. The system performs pre pruning and post pruning to eliminate irrelevant results. The system effectively identifies the Co-Morbid of the diabetes disease and its sub types, the sub type which is referred as the heart disease, retinopathy, neuropathy etc., and the experimental results are evaluated using the Java. The experimental result shows that iterative **SAM** shows better prediction accuracy compared to traditional summarization techniques. Further **FUZZY** is used to predict the comorbid conditions. From the experimental results, the prediction error calculated for diabetes Co-Morbid assessment is almost reduced than the existing system.

## REFERENCES:

- [1] Larose, Daniel T. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.
- [2] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." *ACM SIGMOD Record*. Vol. 22. No. 2. ACM, 1993.
- [3] Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Association rules mining: A recent overview." *GESTS International Transactions on Computer Science and Engineering* 32.1 (2006): 71-82.
- [4] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994.
- [5] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." *ACM SIGMOD Record*. Vol. 29. No. 2. ACM, 2000.
- [6] Centers for Disease Control and Prevention. "National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014." *Atlanta, ga: US Department of health and human services* (2014).
- [7] Diabetes Prevention Program Research Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," *N. Engl. J. Med.*, vol. 346, no. 6, pp. 393-403, Feb. 2002.
- [8] J. Tuomilehto *et al.*, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance," *N. Engl. J. Med.*, vol. 344, no. 18, pp. 1343-1350, May 2001.
- [9] P. W. Wilson, R. B. D'Agostino, H. Parise, L. Sullivan, and J. B. Meigs, "Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus," *Circulation*, vol. 112, no. 20, pp. 3066-3072, Nov. 2005.
- [10] C. Aflori, M. Craus, "Grid implementation of the Apriori algorithm *Advances in Engineering Software*", Volume 38, Issue 5, May 2007, pp. 295-300.
- [11] Srinivas, K., "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", *IEEE Transaction on Computer Science and Education (ICCSE)*, p(1344 - 1349), 2010
- [12] Sa-ngasoongsong, Akkarapol, and Jongsawas Chongwatpol. "An Analysis of DiabetesCo-Morbid Factors Using Data Mining Approach." *Oklahoma state university, USA* (2012).
- [13] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", *IEEE* 2011.
- [14] Burdick, Doug, Manuel Calimlim, and Johannes Gehrke. "MAFIA: A maximal frequent itemset algorithm for transactional databases." *Data Engineering, 2001. Proceedings. 17th International Conference on*. IEEE, 2001.
- [15] V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, Vol. 2, No. 4, 2013, pp 56-66.